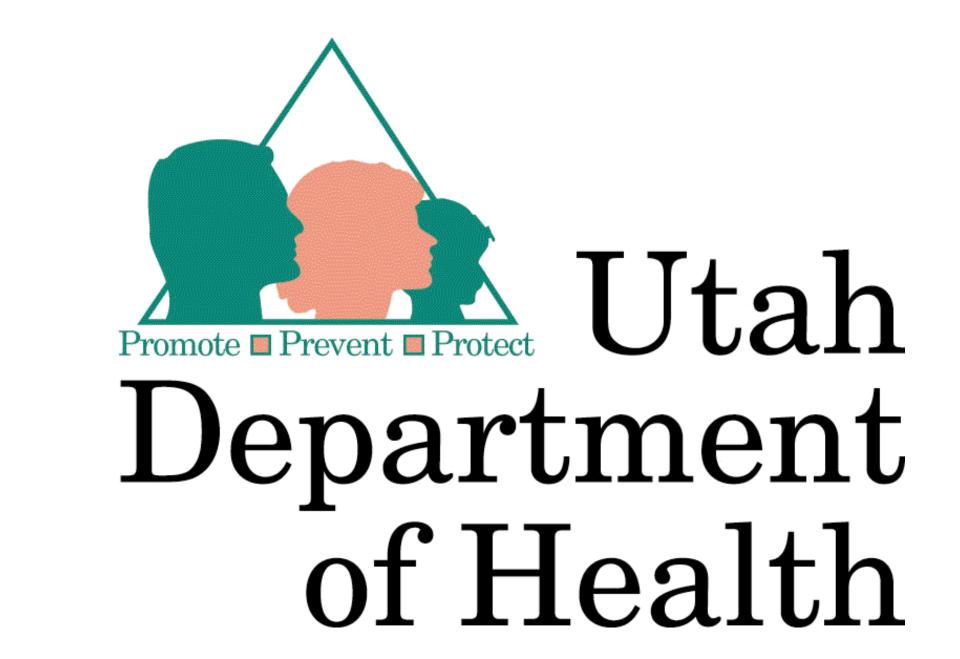# Evidence-based Quality Improvement for Reducing Duplicate Records in the Utah Statewide Immunization Information System (USIIS)

Sandra Schulthies, Yukiko Yoneoka, Chris Pratt, Wu Xu, Ph. D.
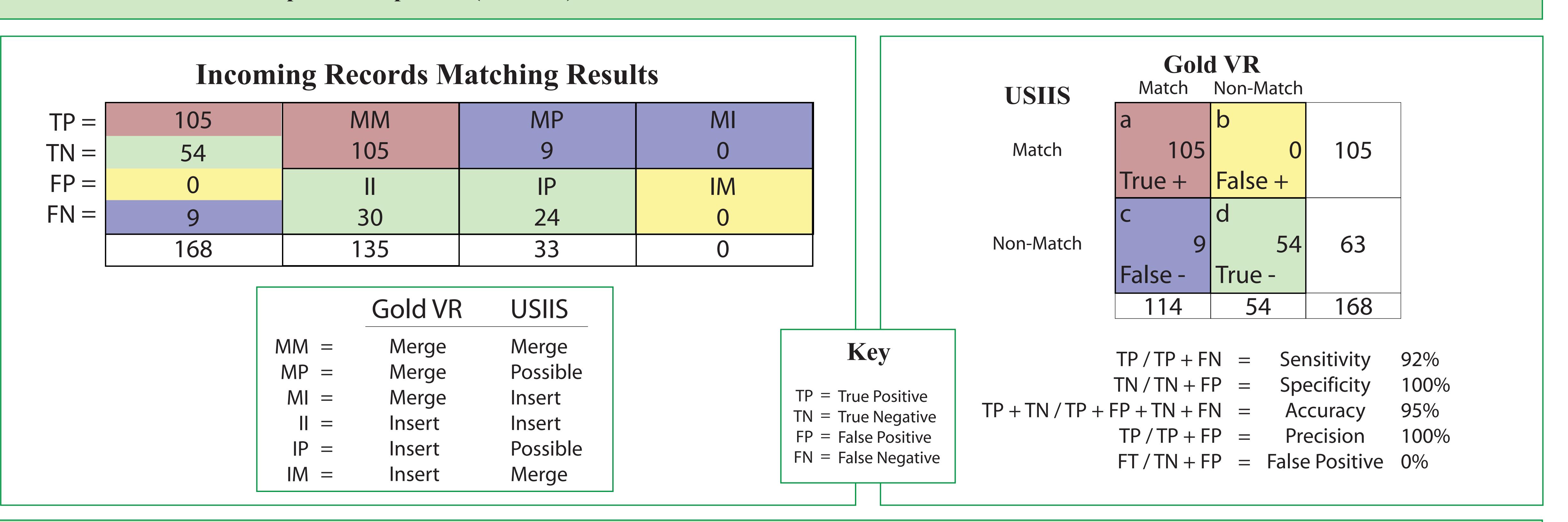*Utah Department of Health, Salt Lake City, Utah*

BACKGROUND

- The Utah Statewide Immunization Information System (USIIS) became fully operational in 1998, and currently contains more than 2.5 million patient records.

- Nearly 450 providers of immunizations submit data to USIIS via flat files or direct data entry into a web application.

- As the database grew, "possible duplicate" patient records created by an inadequate record matching algorithm accumulated to an unmanageable number (more than 400,000), each requiring manual review to resolve.

- A new record matching algorithm and record loading scripts were developed to improve the patient matching process and decrease the number of possible duplicates.

- This study analyzes the effectiveness of the new record matching procedures.

METHODS

- The new USIIS record matching algorithm employs a deterministic and probabilistic approach, using weighted criteria to determine a match. Automated decisions using weighted criteria were based on human de-duplication experience with the unique USIIS data set and Vital Records' birth certificate data.

- 128 stratified-randomly selected incoming records were processed through the new algorithm and matched to existing records resulting in 168 record pairs. Some incoming records matched to more than one existing record, with up to three matches.

- Each patient record pair was manually reviewed by a de-duplication expert using Vital Records' birth certificates to determine a positive or negative match. Two researchers reviewed the match results for the record pairs. Consensus human review results were defined as the "gold standard," for this study and deemed to be the most correct answers. The gold standard was then compared to computer algorithm results for the same records.

- Evaluation metrics were then calculated.

RESULTS

- The new USIIS record matching algorithm classified each record pair into three categories: merge (match – merge with an existing record), possible duplicate (possible match - manual review), or insert (no match – insert as a unique record). The gold standard had only two possible results of either merge/match or insert/non-match.

- Results were tallied for the gold standard and for the USIIS algorithm. Possible duplicate records from the USIIS algorithm were conservatively considered as insert/non-match for this study, which slightly decreased sensitivity.

- Outcomes from the gold standard and USIIS algorithm were formatted into a 2 x 2 table in order to calculate evaluation metrics. Note that the possible duplicates (MP & IP) were included as non-match for USIIS.

Incoming Records Matching Results

TP = 105	MM 105	MP 9	MI 0
TN = 54			
FP = 0	II	IP	IM
FN = 9	30	24	0
168	135	33	0

	Gold VR	USIIS
MM =	Merge	Merge
MP =	Merge	Possible
MI =	Merge	Insert
II =	Insert	Insert
IP =	Insert	Possible
IM =	Insert	Merge

Gold VR

USIIS	Match	Non-Match	
Match	a 105 True +	b 0 False +	105
Non-Match	c 9 False -	d 54 True -	63
	114	54	168

Key

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

TP / TP + FN =	Sensitivity	92%
TN / TN + FP =	Specificity	100%
TP + TN / TP + FP + TN + FN =	Accuracy	95%
TP / TP + FP =	Precision	100%
FT / TN + FP =	False Positive	0%

CONCLUSIONS

- The new matching algorithm exceeds the CDC immunization registry standard of 90% sensitivity and meets the requirement of 100% specificity.

- This same study applied to a sample from the possible duplicates file shows that processing incoming patient records with the new record matching algorithm decreases the number of possible duplicate records by 80%.

- Including human de-duplication experience in the development of automated computer programming decisions was vital to our success.